DOCUMENT RESUME

ED 297 012 TM 011 974

AUTHOR de Gruijter, Dato N. M.
TITLE The Rasch Model and Missing Data, with an Emphasis on Tailoring Test Items.
PUB DATE Apr 88
NOTE 14p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 5-9, 1988).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Adaptive Testing; *Educational Testing; Estimation (Mathematics); Foreign Countries; *Item Analysis; Item Banks; *Test Items
IDENTIFIERS Item Parameters; *Missing Data; Parameter Estimation of Sequential Testing; *Rasch Model

ABSTRACT

Many applications of educational testing have a missing data aspect (MDA). This MDA is perhaps most pronounced in item banking, where each examinee responds to a different subtest of items from a large item pool and where both person and item parameter estimates are needed. The Rasch model is emphasized, and its non-parametric counterpart (the Mokken scale) is considered. The possibility of tailoring test items in combination with their estimation is discussed; however, most methods for the estimation of item parameters are inadequate under tailoring. Without special measures, only marginal maximum likelihood produces adequate item parameter estimates under item tailoring. Fischer's approximate minimum-chi-square method for estimation of item parameters for the Rasch model is discussed, which efficiently produces item parameters. (TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED297012

THE RASCH MODEL AND MISSING DATA, WITH AN EMPHASIS ON TAILORING TEST ITEMS

Paper presented at the AERA conference, New Orleans, April 5-9, 1988

Dato N.M. de Gruijter
University of Leiden

Abstract

Many applications of educational testing have a missing data aspect. This is, for example, the case when person and item parameters are to be estimated in a design in which examinees respond to different tests. Fischer's approximate minimum-chi-square method for the estimation of item parameters for the Rasch model efficiently produces item parameters even in that situation. An exception is the case with a certain amount of tailoring of test items. Without special measures only MML produces adequate item parameter estimates under item tailoring.

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

D.N.M. de Gruijter

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM011974

INTRODUCTION

There are virtually no data without missing data. An example of an application which can be viewed as a missing data problem is the equating or scaling of two different test forms. Here examinees have taken one of two test forms and a possibly internal, common linking test. Responses on one of the two tests are missing. This view on equating within the context of item response theory is nicely exemplified by the approach of Wingersky and Lord (1984), where equating is achieved by estimating all item parameters simultaneously on basis of the available information. The missing data aspect is perhaps most pronounced in item banking, where each examinee responds to a different subtest of items from a large item pool and where both person and item parameter estimates are needed. Such an application will be addressed in the present study.
First, the Rasch model is introduced, and the important and convenient property of this model that the item parameters can be estimated from counts n_{ij}, where n_{ij} is the number of times item i is answered correctly and item j incorrectly. Next a nonparametric counterpart of the Rasch model will be discussed briefly. Then the possibility of tailoring test items in combination with their estimation will be considered. It will be made clear that most methods for the estimation of item parameters are inadequate under tailoring. Next, we will return to the estimation of Rasch item parameters from counts with a discussion of Fischer's (1974) approximate minimum chi-square method, MINCHI for short. Finally, the use of MINCHI in item banking with some amount of tailoring will be treated.

THE RASCH MODEL AND THE MOKKEN SCALE

The probability of a correct response R_i=1 on item i, given ability θ, can be written in the Rasch model as

$$P(R_i=1|\theta) = \exp(\theta-b_i)/[1+\exp(\theta-b_i)], \qquad (1)$$

where b_i is the difficulty parameter of item i. An alternative formulation, which will be used generally in the present context, is

$$P(R_i=1|\xi) = \xi\epsilon_i/(1 + \xi\epsilon_i), \qquad (2)$$

with

$$\xi = \exp(\theta)$$
$$\epsilon_i = \exp(-b_i).$$

Under the assumption of local stochastic independence the probability of a correct response on item i, given one correct response to both item i and j, equals

$$P(R_i=1,\ R_j=0 | R_i+R_j=1) = \epsilon_i/(\epsilon_i+\epsilon_j). \qquad (3)$$

This probability is independent of the ability level ξ. This is a fundamental result for the Rasch model; reversely, the only model which is continuous in terms of its parameters and for which this result holds, is the Rasch model (Fischer, 1974).
Equation 3 implies that the item parameters can be estimated from the relation

$$n_{ij}/(n_{ij} + n_{ji}) \approx p_{ij} = \epsilon_i/(\epsilon_i + \epsilon_j). \qquad (4)$$

The larger the item easiness ϵ_i becomes in comparison to the easiness of item j, ϵ_j, the larger the expected dominance, reflected in Equation 4, becomes. In this respect the Rasch model belongs to a wider class of models, which all share the property of doubly monotone items. Doubly monotone items have item characteristic curves $P(R_i=1|\theta)$ which never cross or touch, except asymptotically. This property implies a weak form of the dominance from Equation 3:

if $P(R_i=1,\ R_j=0 | R_i+R_j=1,\ \theta_0) > 0.5$, for some value θ_0
then $P(R_i=1,\ R_j=0 | R_i+R_j=1,\ \theta\ ) > 0.5$ for all θ.

When no specific assumptions on the form of $P(R_i=1|\theta)$ are made, we have the nonparametric Mokken-scale (Mokken & Lewis, 1982). This model has been implicitly used by Cliff (1977) who proposed the ranking of items on basis of counts n_{ij} in a nonparametric way, starting with the Guttman scale.
Note that ordering of items is trivial in the case of complete data- every examinee responds to all items: ranking on basis of number correct then suffices. The more interesting applications deal with incomplete

data. It should be clear how practical estimation methods based on counts are: as new data are gathered, one only has to update the values n_{ij}.

The models in this section are symmetric with respect to items and persons. So, persons may be ranked on basis of counts s_{ij}, where s_{ij} is the number of times person i answers an item correctly and person does j not do so. This idea will not be pursued, however. With an increasing number of examinees the procedure would become unwieldy. For the Rasch model a simple alternative solution exists: first item parameter estimates may be obtained and secondly person parameter estimates, through maximum likelihood, using the item parameter estimates obtained in the first step.

TAILORING OF TEST ITEMS

We speak of tailoring of test items when the probability that an item will be selected for presentation, depends on the previous responses of the examinee. The purpose of tailored test administration is to present items, adequate for the provisional estimate of the ability level of an examinee. Testing becomes more accurate this way and this enables the test administrator to shorten the test length. In tailored testing it is commonly assumed that accurate item parameter estimates are available. In estimating examinee abilities it then is assumed that the item parameters are known. Cliff (1975) suggested that tailoring might be possible in a nonparametric context even in the absence of a definite rank order of items. Will it be possible to tailor items in the parametric case and at the same time improving initial item parameter estimates?
A simple demonstration will make it clear that a combination of tailoring and estimation fails when estimation of item parameter estimates is based on counts. Table 1 gives the frequencies of response patterns for four Rasch items from which counts n_{ij} can be computed. Evidently, the counts n_{ij} are all equal, implying the equality of item parameters ϵ_i.

Table 1

Now, let us drop the fourth item when the total score on the first three items equal 2 or 3. This simulates tailoring where the fourth item is administered only after a total score less then two on the other items. In this case $n_{i4} < n_{4i}$ $(i = 1, 2, 3)$, i.e. the fourth item seems easier

than the other items. This effect was demonstrated by Fischer (1986). The bias will disappear under special conditions, when items do not discriminate or discriminate perfectly, i.e. in case of the Guttman scale. Cliff (1975, 1977), who suggested tailoring in combination with ordering the items for imperfect Guttman scales, mentioned the possibility of an artificial degree of consistency in his approach.

The bias was also discussed by De Gruijter (1980), in connection with ML. He argued that the effect arises because of the imperfect measurement in a routing test. Another way to describe the effect is in terms of ignorability (Rubin, 1976). The local independence, formulated as

$$P(R_1=r_1,\ldots,R_m=r_m) = \prod_{i=1}^{m} P(R_i=r_i), \qquad (5)$$

does not describe the process accurately with missing data in tailoring when item parameters are to be estimated. However, under marginal maximum likelihood (MML) the missing data process can be ignored. De Gruijter (1980) argued that characteristics of the posterior distribution after a routing test can be used to obtain unbiased item parameter estimates. The robustness of MML was proved mathematically for the Rasch model by Glas (1988).

There is one additional problem to which some attention should be given. In the Rasch model the response data are condensed in the form of the sufficient statistic, the total score. In MML this implies the introduction of an elementary symmetric function, which relates to all possibilities to obtain the total score. In tailoring test items there will be a smaller number of combinations leading to a given total score on basis of a given subset of items than the case of no tailoring. Take for example two-stage testing. The total score on the test t can be written as t_1+t_2, where t_1 is the total score on the routing test and t_2 is the total score on a particular second-stage test. In order to have this second-stage test t_1 should lie within a certain interval and this restricts the number of ways in which the total score t can be obtained. Fortunately, the elementary symmetric functions can be eliminated in the estimation process (Thissen, 1982).

So the tailoring process may be ignored completely under MML and item parameter estimates can be obtained in the same way as when item responses are missing randomly. This implies that estimation can be done with any MML-estimation program which allows for missing data. The unbiasedness

of MML presents a strong case for its use in applications. However, MML has disadvantages too. In practice data may be gathered over a long period of time. During this period the composition of the examinee population might change. When at the same time the composition of the item pool changes through the addition of new items, MML on basis of the assumption of one fixed population will fail (De Gruijter, 1987). For this reason it is important to investigate whether bias in other estimation methods might be circumvented. It will become clear that this is possible with a method based on counts under partial tailoring conditions. First, we will introduce the MINCHI-procedure in the next section.

MINCHI

Fischer (1974) suggested an approximate minimum chi-square procedure based on Equation 3. More specifically, he suggested to minimize the function

$$F = \sum_{i<j} \sum [n_{ij} - (n_{ij}+n_{ji})p_{ij}]^2 / [(n_{ij}+n_{ji})p_{ij}(1-p_{ij})] \qquad (6)$$

w.r.t. the parameters. Differentiating and setting the results equal to zero gives a set of equations

$$\epsilon_i^{-2} = \sum_j x_{ji} \epsilon_j^{-1} / (\sum_j x_{ij} \epsilon_j), \qquad (i=1,\ldots,m) \qquad (7)$$

with

$$x_{ij} = n_{ij}^2 / (n_{ij}+n_{ji})$$

and m items in total. This set of equations can be solved very fast iteratively. Results from simulation studies (Fischer, 1974; Zwinderman & Van den Wollenberg, 1987) indicate that the method is accurate as well. This might be a bit surprising while the method is only an approximation to minimum chi-square. Dependencies between counts n_{ij} and n_{ik} are neglected in the target function F from Equation 6. An alternative would be to eliminate the dependencies, which Van der Linden and Eggen (1986) did by selectively removing data for each examinee. The impact

of the dependencies should decrease when different examinees obtain different subtests from a large item pool.

A graphical model check based on counts n_{ij} was already suggested by Rasch (1960).This suggestion will be slightly modified in connection with estimation using MINCHI. Given counts n_{ij} and n_{ji} greater than zero, the difference between b_j and b_i (Equation 1) is estimated as

$$\text{est}(b_j - b_i) = \ln (n_{ij}/n_{ji}). \qquad (8)$$

The differences can be computed for fixed i, with $\text{est}(b_i - b_i)=0$, and plotted against the MINCHI-estimates $\hat{b}_j$. When the model fits, the points should lie along a straight line, with a slope of 45 degrees and an intercept d_i equal to b_i.

For purposes of illustration a simulation was done with a 'Rasch' model with a common guessing parameter equal to 0.25. Ten equally spaced bs ranged from -2.0 to 2.5. The simulation entailed the computation of relative frequencies f_{ij}, instead of counts, for an infinite population, according to a rough approximation to the standard normal distribution. For each item a plot was made as suggested above. The results are given in Figure 1. From this figure no violations of the Rasch

Figure 1

model can be detected. Clearly some model deviations, like those of the example with double-homogeneous items, cannot be detected with counts n_{ij}. There is an obvious reason for this insensitivity. The counts are obtained by using weighted averages over score groups (assuming complete data) and with this averaging information is lost.

The results demonstrate that one should preferably check the extent of model fit before condensing the data into counts n_{ij} (De Gruijter, 1987). When the items are double-homogeneous MINCHI might very well be able to rank the items appropriately, due to the fact that these items exhibit a weak form of the property in Equation 3. In this case however, the exact parameters should not be taken seriously.

MINCHI-ESTIMATION AND TAILORING

As long as there is no tailoring, MINCHI seems a good method for the estimation of Rasch item parameters, especially when the data are incomplete. The results of the section on tailoring indicate that MINCHI cannot be used under a full-fledged tailoring approach to testing. Fortunately, this does not mean that the estimation of item parameters or the updating of item parameters becomes impossible with even the slightest amount of tailoring. On the contrary, as long as tailoring is not complete, estimation remains possible. This can be illustrated with the presentation schemes in Figure 2.

Figure 2

In both schemes we have two-stage testing. The second-stage test is tailored to the estimated ability level of examinees after a routing test of n_1 items. In the first scheme items for the routing test are selected from a fixed subset A of the item pool. When the score on the routing test exceeds t, n_2 items are selected from subset B, otherwise n_2 items are selected from subset C.

The items in A are not selected on basis of previous responses. So the use of counts n_{ij} in order to obtain item parameters is legitimate for items in subset A. The responses to items in B (or C) depend only on the unknown abilities θ, and item parameter estimation based on counts n_{ij} is also possible for B(or C). It is illegitimate, however, to use counts n_{ij} for $i(j)$ in A and $j(i)$ in B or C. So, three separate Rasch scales are obtained.

One common Rasch scale can be obtained when the presentation scheme is slightly changed like in the righthand side of Figure 2. In this case each item is eligible for the routing test and information on item pairs with one item in A and the other item in B (or C) becomes available. To be more specific, assume $i \in$A, $j \in$B. Then it is possible to update count n_{ij} or n_{ji} when both items i and j are selected for the routing test. If both are selected, but j is selected for the second-stage test, updating is not allowed. This scheme is, of course, only one out of the multitude of possibilities of partial tailoring where a common Rasch scale is obtained. A sound procedure would be to have little or no tailoring in

the first phase of data collection, and to increase the degree of tailoring when increasing the accuracy of item parameter estimates becomes less important.

DISCUSSION

In many applications one wants to have information on examinees as efficiently as possible. Tailoring of test items is a tool in order to increase testing efficiency. Usually tailoring is done with parameters assumed to be known. Tailoring and item parameter estimation are not incompatible under MML. Other estimation methods will have to be adapted in order to make partial tailoring and estimation possible. For the Rasch model there is a very simple estimation method for item parameters based on counts, which can easily be adapted to a situation where some of the items are administered on basis of previous examinee responses. The adaptation simply consists of ignoring data influenced by the tailoring process.

References

Cliff, N. (1975). Complete orders from incomplete data: interactive ordering and tailored testing. *Psychological Bulletin, 82*, 289-302.

Cliff, N. (1977). A theory of consistency of ordering generalizable to tailored testing. *Psychometrika, 42*, 375-399.

De Gruijter, D.N.M. (1980). A two-stage testing procedure. In L.J.Th. van der Kamp, W.F. Langerak & D.N.M. de Gruijter (Eds.), *Psychometrics for educational debates*. Chichester: Wiley.

De Gruijter, D.N.M. (1987). On the robustness of the "Minimum-Chi-Square" method for the Rasch model. *Tijdschrift voor Onderwijsresearch, 12*, 225-232.

Fischer, G.H. (1974). *Einführung in die Theorie psychologischer Tests*. Bern: Huber.

Fischer, G.H. (1986). Personal communication.

Glas, C.A.W. (1988). The Rasch model and multistage testing. *Journal of Educational Statistics, 13*, 45-52.

Mokken, R.J., & Lewis, Ch. (1982). A nonparametric approach to the analysis of dichotomous item responses. *Applied Psychological Measurement, 6*, 417-430.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Rubin, D.B. (1976).Inference and missing data. *Biometrika, 63*, 581-592.

Thissen, D. (1982). Marginal maximum likelihood estimation for the one-parameter logistic model. *Psychometrika, 47*, 175-186.

Van der Linden, W.J., & Eggen, T.J.H.M. (1986). An empirical Bayesian approach to item banking. *Applied Psychological Measurement, 10*, 345-354.

Wingersky, M.S., & Lord, F.M. (1984). An investigation of methods for reducing sampling error in certain IRT procedures. *Applied Psychological Measurement, 8*, 347-364.

Zwinderman, K.H., & Van den Wollenberg, A.L. (1987). Minimum chi-square estimation in the Rasch model. Paper presented at the European Meeting of the Psychometric Society.

Table 1. Frequency distribution for response patterns with four items.

pattern	frequency (N large)
1110	$9N$
1101	$9N$
1011	$9N$
0111	$9N$
1100	$3N$
1010	$3N$
0110	$3N$
1001	$3N$
0101	$3N$
0011	$3N$
1000	$1N$
0100	$1N$
0010	$1N$
0001	$1N$

$A \cap B = \phi$, $A \cap C = \phi$, $B \cap C = \phi$

n_1 items $i \in A$

$t = \sum_j^{n_1} r_j$

if $t > t_c$ n_2 items $i \in B$

else n_2 items $i \in C$

(a) two-stage testing with test forms selected from disjoint item sets A, B en C

$A \cap B = \phi$, $A \cap C = \phi$, $B \cap C = \phi$

n_1 items $i \in A \cup B \cup C$

$t = \sum_j^{n_1} r_j$

if $t > t_c$ n_2 items $i \in B \cap \tilde{T}_1$

else n_2 $i \in C \cap \tilde{T}_1$

(b) modified two-stage testing procedure; T_1 is the set of items selected for the routing test

Figure 2. Two possible item presentation schemes with partial item tailoring.

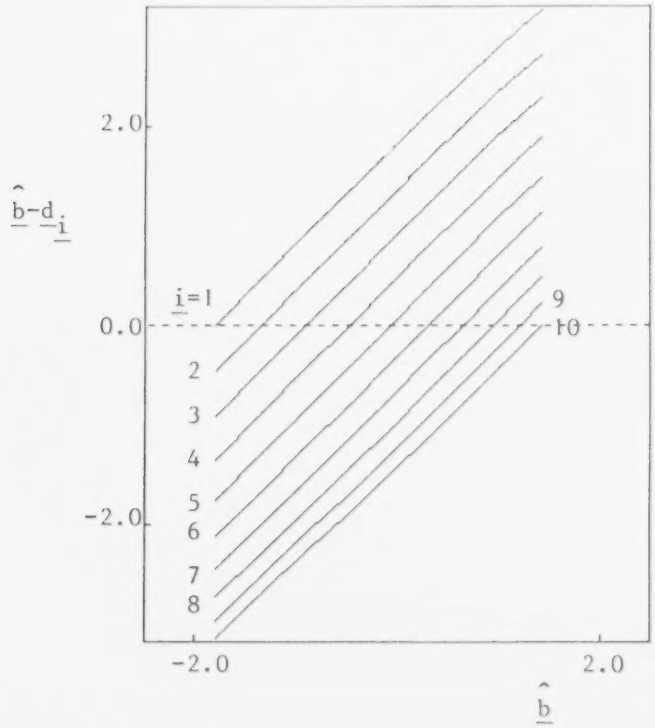

Figure 1. Estimates $b_j - b_i = \ln (n_{ji}/n_{ij})$ for fixed i plotted against MINCHI-estimates $\hat{b}_j$.